

Efficient Pattern Matching in Elastic-Degenerate Strings^{*}

Costas S. Iliopoulos, Ritu Kundu, and Solon P. Pissis

Department of Informatics, King's College London, London WC2R 2LS, UK
 {costas.iliopoulos, ritu.kundu, solon.pissis}@kcl.ac.uk

Abstract. In this paper, we extend the notion of gapped strings to *elastic-degenerate strings*. An elastic-degenerate string can be seen as an ordered collection of $k > 1$ seeds (substrings/subpatterns) interleaved by *elastic-degenerate symbols* such that each elastic-degenerate symbol corresponds to a set of two or more variable length strings.

Here, we present an algorithm for solving the pattern matching problem with (solid) pattern and elastic-degenerate text, running in $\mathcal{O}(N + \alpha\gamma nm)$ time; where m is the length of the given pattern; n and N are the length and total size of the given elastic-degenerate text, respectively; α and γ are small constants, respectively representing the maximum number of strings in any elastic-degenerate symbol of the text and the largest number of elastic-degenerate symbols spanned by any occurrence of the pattern in the text. The space used by the algorithm is linear in the size of the input for a constant number of elastic-degenerate symbols in the text; α and γ are so small in real applications that the algorithm is expected to work very efficiently in practice.

Keywords: pattern matching, elastic-degenerate strings, degenerate strings, indeterminate strings, gapped patterns

1 Introduction

Uncertainty in sequential data (strings) can be characterised using various representations. One such representation is *degenerate string* which is defined by the existence of one or more positions that are represented by sets of symbols from an alphabet Σ , unlike an accurate or certain (standard) string characterised by a single symbol at each position. For instance, $[\text{a}]_1 \text{ac} [\text{c}]_2 \text{a} [\text{b}]_3$ is a degenerate string of length 6 over $\Sigma = \{\text{a}, \text{b}, \text{c}\}$.

A *compound pattern* (or *gapped pattern*) is another way to capture uncertainty – it is a list of standard (simple) sub-patterns (or seeds) separated by variable length gaps defined by a list of intervals [5]. Simply, a compound pattern P can be represented as follows [15]: $P = P_1 *^{a_1, b_1} P_2 *^{a_2, b_2} P_3 \dots *^{a_{l-1}, b_{l-1}} P_l$ where, $*$ is a *wildcard character* (also called *don't care symbol* or *hole*) that matches any character in a finite alphabet Σ ; $\forall i \in [1..l]$ each sub-pattern P_i is a string over Σ ;

^{*} This work was partially supported by the British Council funded INSPIRE Project

and $\forall i \in [1..l-1]$ each pair (a_i, b_i) represents the gap (minimum and maximum wildcard characters, respectively) between two consecutive subpatterns P_i and P_{i+1} .

Here, we introduce another representation to encapsulate uncertainty in sequential data – which we call *elastic-degenerate strings* – by extending and combining the idea of gapped patterns/strings and degenerate strings. An *elastic-degenerate string* is a string such that at one or more positions, an *elastic-degenerate symbol* can occur which is defined as a set of variable length substrings. Another way to visualise an elastic-degenerate string is to see it as an ordered collection of $k > 1$ seeds (substrings) interleaved by elastic-degenerate symbols such that each elastic-degenerate symbol corresponds to a set of two or more variable length substrings. $bc \begin{bmatrix} ab \\ aab \\ aca \end{bmatrix} ca \begin{bmatrix} abcab \\ cba \end{bmatrix} bb$ is an example of an elastic-degenerate string over $\Sigma = \{a, b, c\}$.

This generalisation of concept of *degeneracy* is motivated by several important data mining problems which can be reduced to the core task of discovering occurrences of one or more patterns in a text that can best be described as an ordered collection of strings interleaved by sets of variable length strings.

More specifically, in genomics an important class of problems is to study within-species genetic variation; the state of the art solutions for this class comprises of matching(*mapping*) substrings (called *reads*) to a longer genomic sequence (canonical *reference genome* obtained through assembly). Owing to the high diversity among biologically relevant genomic regions in many organisms, the population level complexities can not be captured by the ‘linear’ structure of a reference genome (see [11]). Consequently, recent research trend has shifted towards using alternative representations of genomic sequence for population-based genome assembly ([9,2,6,12]). One such representation that encodes a set of related genomes with variations in the reference genome itself (called Population Reference Genome in [12]), can be seen as an elastic-degenerate string.

The problem of matching in the context of gapped strings has been studied extensively using combinatorial approaches (see [14] and references therein). However, a gapped string (which specifies the constraint on only the length of the gap between two consecutive seeds) differs from an elastic-degenerate string because the later precisely defines the possible substrings (of varying lengths) that can exist between those consecutive seeds. This precise identification of allowed substrings in a gap makes the matching problem in the context of elastic-degenerate strings, algorithmically more challenging and computationally difficult.

In this paper, we not only formalize the concept of elastic-degenerate strings but also present an efficient - in terms of both, space and time - algorithm to solve the pattern matching problem in a given elastic-degenerate text. To the best of our knowledge, no other work, heretofore explores the problem accounting for *elastic-degeneracy* in the text.

In the next section, we introduce the basic definitions and establish the notions of elastic-degeneracy that will be used in this paper. The algorithmic tools

required to build the solution are described in Section 3. In Section 4, we formally define the problem along with presenting the algorithm. The algorithm is analysed in Section 5. Finally, the paper is concluded in Section 6.

2 Terminology and Technical Background

We begin with basic definitions and notations. We think of a *string* X of *length* n as an array $X[1..n]$, where every $X[i]$, $1 \leq i \leq n$, is a *letter* drawn from some fixed *alphabet* Σ of size $|\Sigma| = \mathcal{O}(1)$. The *empty string* is denoted by ε . Σ^* denotes the set of all strings over an alphabet Σ including empty string ε . A string Y is a *factor* of a string X if there exist two strings U and V , such that $X = UYV$. Hence, we say that there is an *occurrence* of Y in X , or simply, that Y *occurs in* X . The starting position of an occurrence, say i , is called *head* of the occurrence and its ending position ($i + |Y| - 1$) is called its *tail*. Note that an empty string occurs at each position in a given string.

Consider the strings X , Y , U , and V , such that $X = UYV$. If $U = \varepsilon$, then Y is a *prefix* of X . If $V = \varepsilon$, then Y is a *suffix* of X .

A *degenerate symbol* $\tilde{\sigma}$ over an alphabet Σ is a non-empty subset of Σ , i.e., $\tilde{\sigma} \subseteq \Sigma$ and $\tilde{\sigma} \neq \emptyset$. $|\tilde{\sigma}|$ denotes the size of the set and we have $1 \leq |\tilde{\sigma}| \leq |\Sigma|$. A *degenerate string* is built over the potential $2^{|\Sigma|} - 1$ non-empty sets of letters belonging to Σ . In other words, a degenerate string $\tilde{X} = \tilde{X}[1..n]$, is a string such that every $\tilde{X}[i]$ is a degenerate symbol, $1 \leq i \leq n$. If $|\tilde{x}[i]| = 1$, that is, $\tilde{X}[i]$ represents a single symbol of Σ , we say that $\tilde{X}[i]$ is a *solid symbol* and i is a *solid position*. Otherwise $\tilde{X}[i]$ and i are said to be a *non-solid symbol* and a *non-solid position*, respectively. For example, $[\mathbf{a}]_{\mathbf{ac}}[\mathbf{b}]_{\mathbf{a}}[\mathbf{c}]_{\mathbf{c}}$ is a degenerate string of length 6 over $\Sigma = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. A string containing only solid symbols will be called a *solid string*. A *conserved degenerate string* is a degenerate string where its number of non-solid symbols is upper-bounded by a fixed positive constant k .

Now we give the terminology to build the concept of elastic-degeneracy by presenting the following definitions and examples.

Definition 1 (Seed: S). A seed S is a (possibly empty) string over Σ (i.e. $S \in \Sigma^*$).

Definition 2 (Elastic-Degenerate Symbol: ξ). An elastic-degenerate symbol ξ , over a given alphabet Σ , is a non-empty set of strings over Σ (i.e. $\xi \subset \Sigma^*$ and

$\xi \neq \emptyset$). An elastic-degenerate symbol ξ is denoted by $\begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{|\xi|} \end{bmatrix}$, where each E_i , $1 \leq$

$i \leq |\xi|$ is a solid string. The minimum (maximum) length of ξ , represented as $|\xi|_{\min}$ ($|\xi|_{\max}$), is the length of the shortest (or longest) string in the set.

Definition 3 (Elastic-Degenerate String: \hat{X}). An elastic-degenerate string \hat{X} , over a given alphabet Σ , is a sequence $S_1\xi_1S_2\xi_2S_3..S_{k-1}\xi_{k-1}S_k$, where S_i , $1 \leq i \leq k$ is a seed and ξ_i , $1 \leq i \leq k - 1$ is an elastic-degenerate symbol.

An elastic degenerate string \hat{X} can be visualised as follows:

$$\hat{X} = S_1 \begin{bmatrix} E_{1,1} \\ E_{1,2} \\ \vdots \\ E_{1,|\xi_1|} \end{bmatrix} S_2 \begin{bmatrix} E_{2,1} \\ E_{2,2} \\ \vdots \\ E_{2,|\xi_2|} \end{bmatrix} S_3 \dots S_{k-1} \begin{bmatrix} E_{k-1,1} \\ E_{k-1,2} \\ \vdots \\ E_{k-1,|\xi_{k-1}|} \end{bmatrix} S_k$$

Example 1 $\hat{X} = abbcc \begin{bmatrix} ab \\ aab \\ acca \end{bmatrix} cca \begin{bmatrix} aabcab \\ cba \end{bmatrix} bb$ is an elastic-degenerate string, where we have the following:

- Three seeds: $S_1 = abbcc$, $S_2 = cca$, and $S_3 = bb$.
- Two elastic-degenerate symbols:
 $\xi_1 = \begin{bmatrix} ab \\ aab \\ acca \end{bmatrix}$ and $\xi_2 = \begin{bmatrix} aabcab \\ cba \end{bmatrix}$.
- For ξ_1 : $E_{1,1} = ab$, $E_{1,2} = aab$, $E_{1,3} = acca$; minimum length is 2 (length of $E_{1,1}$ is shortest); and maximum length is 4 (length of $E_{1,3}$ is longest).
- For ξ_2 : $E_{2,1} = aabcab$, $E_{2,2} = cba$; minimum length is 3 (length of $E_{2,1}$ is shortest); and maximum length is 6 (length of $E_{2,1}$ is longest).

Observe the use of \hat{X} to distinguish an elastic degenerate string from a (plain) solid string X or a degenerate string \tilde{X} . In the following, we define three characteristics of a given elastic degenerate string \hat{X} (with k seeds).

Definition 4 (Total Size: $\|\hat{X}\|$). Total size of \hat{X} , represented as $\|\hat{X}\|$, is defined as the sum of the total length of its seeds and the total length of all the strings in each of its elastic-degenerate symbols (i.e. $\|\hat{X}\| = \sum_{i=1}^k |S_i| + \sum_{i=1}^k \sum_{j=1}^{|\xi_i|} |E_{i,j}|$).

Definition 5 (Length: $|\hat{X}|$). The length of \hat{X} is denoted by $|\hat{X}|$ and is defined as the sum of the total length of its seeds and the total number of its elastic-degenerate symbols (i.e. $|\hat{X}| = \sum_{i=1}^k |S_i| + k - 1$. Informally, the total number of positions in \hat{X} is its length (considering an elastic-degenerate symbol to occupy only one position). Intuitively, a position belonging to some seed will be called solid position and that of an elastic-degenerate symbol will be called elastic-degenerate position.

In the running example, the total length of the seeds is 9; hence, $\|\hat{X}\| = 9 + (2 + 3 + 4) + (6 + 3) = 27$, while $|\hat{X}| = 9 + 2 = 11$. The first **a** occurs at (solid) position 1, followed by **b** at (solid) position 2 and so on; ξ_1 and ξ_2 are at (elastic-degenerate positions) 4 and 9, respectively; the last **b** is at (solid) position 11.

Definition 6 (Possibility-Set: \mathfrak{R}). The possibility-set \mathfrak{R} of \hat{X} is a set of all possible (plain) solid strings obtained from \hat{X} . A solid string can be obtained by replacing each of the elastic-degenerate symbols with one of its constituent strings. More formally, it can be defined as follows:

$$\mathfrak{R} = \{S_1 E_{1,r_1} S_2 E_{2,r_2} \dots E_{k-1,r_{k-1}} S_k\} \quad \forall r_i, 1 \leq i \leq k-1 \text{ such that } 1 \leq r_i \leq |\xi_i|$$

For instance, in the running example, $\mathfrak{R} = \{\text{abbcabccaaabcbabb}, \text{abbcabccacbabbb}, \text{abbcabcccaabcbabb}, \text{abbaabccacbabbb}, \text{abbcaccaccaabcbabb}, \text{abbcaccaccacbabbb}\}$ (constituent strings replacing the elastic-degenerate symbols have been underlined for clarity).

Now we can define *matching* and *occurrence* in the context of elastic-degenerate strings.

Definition 7 (Matching). A given elastic-degenerate string \hat{X} is said to match a solid string Y if, and only if, $Y \in \mathfrak{R}$ of \hat{X} . It is represented as $\hat{X} \simeq Y$. Informally, if one of the possible strings obtained from \hat{X} is the same as Y , we say that they match.

Analogously, two given elastic-degenerate strings \hat{X} and \hat{Y} match (represented as $\hat{X} \simeq \hat{Y}$) if and only if $\hat{X} \cap \hat{Y} \neq \emptyset$. Stating informally, both \hat{X} and \hat{Y} must produce at least one common solid string.

Example 2 Consider \hat{X} as given in Example 1. If a string $Y = \text{abbcabccacbabbb}$ then $\hat{X} \simeq Y$ whereas for a string $Z = \text{abcccccca}$, $\hat{X} \not\simeq Z$ as Z does not occur in the possibility-set \mathfrak{R} of \hat{X} . Given an elastic-degenerate string $\hat{Y} = \text{ab} \begin{bmatrix} \text{bcab} \\ \text{abb} \end{bmatrix} \text{ccacbabbb}$, $\hat{X} \simeq \hat{Y}$ as abbcabccacbabbb is a common solid string obtained from both.

Definition 8 (Occurrence). Given two positions i and j in an elastic-degenerate string (text) \hat{T} , let S be some solid string obtained from $\hat{T}[i..j]$ (i.e. $S \in \mathfrak{R}$). A given solid string (pattern) P is said to occur in \hat{T} between positions i and j , if

$$\begin{cases} P = S & \text{if both, } i \text{ and } j, \text{ are solid} \\ P \text{ is prefix of } S & \text{if } i \text{ is solid and } j \text{ is elastic-degenerate} \\ P \text{ is suffix of } S & \text{if } i \text{ is elastic-degenerate and } j \text{ is solid} \\ P \text{ is factor of } S & \text{if both, } i \text{ and } j, \text{ are elastic-degenerate} \end{cases}$$

An occurrence is represented as the pair of start-position (head) and end-position (tail).

For consistency with the intuitive meaning of an occurrence, we say that P occurs at the position of some elastic-degenerate symbol (say ξ_i) of \hat{T} , if it is a factor of any of the constituent strings of ξ_i .

Example 3 Consider a pattern $P = \text{cabbcb}$ and a text \hat{T} as follows:

$$\text{aacabbbcbbc} \begin{bmatrix} a \\ \text{aab} \\ \text{acca} \end{bmatrix} \text{bb} \begin{bmatrix} c \\ \text{acabbbcb} \\ \text{cba} \end{bmatrix} \text{bacabbc} \begin{bmatrix} b \\ \text{cabb} \\ \text{bbc} \\ \text{aacabb} \end{bmatrix} \text{cbc}$$

All the occurrences of P in \hat{T} are given in Table 1.

Note that more than one occurrence of P can start from the same position but their ending-positions are different (for instance, (11, 14) and (11, 15) in Example 3). Also, note that different strings in the same elastic-degenerate symbols can lead to the same occurrence i.e. same pair of head and tail (as happened for occurrences (17, 22) and (17, 24) in Example 3).

Occurrence:	(3, 8)	(10,15)	(11,14)	(11,15)	(14,14)	(17,22)	(22,24)
Strings chosen:	-	$\xi_1: \underline{a}$ $\xi_2: \underline{c}$	$\xi_1: \underline{acga}$ $\xi_2: \underline{cba}$	$\xi_1: \underline{acga}$ $\xi_2: \underline{c}$	$\xi_2: \underline{acgabbcb}$	$\xi_3: \underline{c}$ or $\xi_3: \underline{cbc}$	$\xi_3: \underline{cabb}$ or $\xi_3: \underline{aac}$

Table 1. Table representing the occurrences of P in \hat{T} as given in Example 3.

Example 4 Here, we illustrate the case, where an elastic-degenerate string has an empty string as a seed. Consider

$$\hat{T} = ab \begin{bmatrix} bcab \\ abb \end{bmatrix} \begin{bmatrix} ab \\ cbb \\ abc \end{bmatrix} cca \begin{bmatrix} bb \\ cb \end{bmatrix} ca \text{ and a pattern } P = babbcb,$$

there is an occurrence of P at $(2, 4)$ of \hat{T} .

3 Algorithmic Tools

Here, we briefly introduce a fundamental data structure, which supports a wide variety of string matching algorithms, and a well-known pattern matching algorithm. This data structure and pattern matching algorithm will be used by the proposed algorithm.

Suffix Tree

The *suffix tree* $\mathcal{S}(X)$ of a non-empty string X of length n , is a compact trie representing all the suffixes of X such that $\mathcal{S}(X)$ has n leaves, labelled from 1 to n . For a general introduction to suffix trees, see [3]. The construction of the suffix tree $\mathcal{S}(X)$ of the input string X takes $\mathcal{O}(n)$ time and space, for string over a fixed-sized alphabet [18,13,17]. Once the suffix tree of a given string (called text) has been constructed, it can be used to support queries that return the occurrences of a given string (called pattern) in time linear in the length of the pattern. Least Common Ancestor (LCA) of the two leaves of a suffix tree can be computed in constant time after a linear time preprocessing to answer LCA queries [8,16]. A generalised suffix tree is a suffix tree for a set of strings [1,7].

KMP Algorithm and failure function

Knuth, Morris and Pratt (KMP) discovered the first linear time string-matching algorithm [10], that is the problem of finding all occurrences of a pattern P in a text T . The KMP algorithm follows the naïve approach for this problem, that is, it slides the pattern across the text. Additionally it preprocesses the pattern P by computing a *failure function* f that indicates the largest possible shift, using previously performed comparisons. Specifically, the *failure function* $f(i)$ is defined as the length of the longest prefix of P that is a suffix of $P[1..i]$. By using the failure function, it achieves an optimal search time of $\mathcal{O}(n)$ after $\mathcal{O}(m)$ -time pre-processing, where n is the length of T and $m < n$ is the length of P .

4 Algorithm for pattern matching in elastic-degenerate text

4.1 Problem Definition

PROBLEM: FINDING OCCURRENCES IN ELASTIC-DEGENERATE TEXT GIVEN A SOLID PATTERN

Input: A pattern P of length m , an elastic-degenerate text $\hat{T} = S_1\xi_1S_2\ldots\xi_{k-1}S_k$, of length n and total size N , where each $\xi_i = \{E_{i,j}\}$, $1 \leq j \leq |\xi_i|$.

Output: All the occurrences of P in \hat{T} .

All the occurrences of the pattern P in the text \hat{T} , fall under the following cases:

1. P entirely lies in some seed S_i .
2. P entirely lies in some string of an elastic-degenerate symbol ξ_i .
3. P spans across one or more elastic-degenerate symbols. This can further be seen as:
 - (a) P starts in some seed S_i .
 - (b) P begins in some string of an elastic-degenerate symbol ξ_i .

For instance, consider Example 3 - the occurrences (3, 8) and (14, 14) lie in Case 1 and Case 2, respectively; (10, 15) and (17, 22) belong to Case 3(a); Case 3(b) covers (11, 14), (11, 15), and (22, 24).

4.2 Algorithm

We now present an efficient algorithm that makes use of the KMP pattern matching algorithm and the suffix tree. Clearly, KMP pattern matching algorithm can easily report the occurrences corresponding to the Cases 1 and 2. Case 3 requires some additional processing and data-structures. The algorithm works in two stages, outlined in the following:

Stage 1: Pre-processing Preprocess the pattern P to compute its failure-function as required for KMP algorithm. In addition, create a generalised suffix tree \mathcal{ST}_{S_i} for the set $\{P, S_i\}$ corresponding to each seed S_i , $1 \leq i \leq k$, as well as a generalised suffix tree \mathcal{ST}_{ξ_i} for the set $\{P, E_{i,1}, E_{i,2}, \ldots, E_{i,|\xi_i|}\}$ corresponding to each elastic-degenerate symbol ξ_i , $1 \leq i \leq k - 1$. Furthermore, pre-process these suffix trees so as to answer the longest common ancestor (LCA) queries in constant time.

Stage 2: Search Start searching the pattern in the text using the KMP algorithm, comparing the letters and using failure function to shift the pattern on a mismatch. The starting position of an occurrence being tested may be either solid or elastic-degenerate; we call the two types of occurrences as *Type 1* and *Type 2*, respectively. We consider the two types separately as follows:

Type 1: Solid start-position Consider a situation, where an occurrence starting from a position (say pos) that lies in some seed S_i , is being tested. Proceed normally comparing the corresponding letters of P and S_i ; shifting the pattern using failure function on mismatch. As soon as the elastic-degenerate symbol ξ_i is encountered (suppose corresponding position in the pattern is p), abort the KMP algorithm (for this test). Check each of the strings of ξ_i (i.e. $E_{i,j}$) whether or not it occurs in the pattern at position p , using LCA queries on $\mathcal{ST}(\xi_i)$; *ticking* (marking) the tails of the found occurrences. It can be realized by maintaining a boolean array of size m , called \mathcal{TT}_i .

Next, Procedure 1 (given below) is followed; in which each ticked position of \mathcal{TT}_i is tried to extend by testing whether S_{i+1} occurs adjacent to it (using LCA queries on $\mathcal{ST}_{S_{i+1}}$). For each such found occurrence of S_{i+1} , occurrences of strings of ξ_{i+1} are checked using the suffix tree $\mathcal{ST}_{\xi_{i+1}}$ and their tails are ticked in \mathcal{TT}_{i+1} . The procedure will then be repeated for \mathcal{TT}_{i+1} ; it continues recursively until there is no tail marked in some call.

Once the process ends (reporting all the occurrences of P starting from pos , if any), the failure function corresponding to the position where the KMP algorithm was aborted (i.e. p) is used to shift the pattern and the KMP algorithm resumes.

It is to be noted that an occurrence of P is implied, if the length of LCA of the pattern starting from some ticked-tail t with either of the following hits the boundary of the pattern:

- some seed S_j (i.e. $|LCA_{t,S_j}| + t > m$)
- any string $E_{i,j}$ of some elastic-degenerate symbol ξ_i (i.e. $|LCA_{t,E_{i,j}}| + t > m$).

Figure 1 elucidates the description given above.

Type 2: Elastic-Degenerate start-position Now consider a situation, where the starting position of an occurrence to be tested is an elastic-degenerate symbol ξ_i . This case can be processed in the similar fashion as one described for Type 1, with the only difference in the manner in which tails are ticked initially.

Begin by applying the KMP algorithm for each $E_{i,j}$, to achieve two purposes: finding the occurrences of P in $E_{i,j}$ and ticking the last position of $E_{i,j}$ for which a prefix of P appears as a suffix of $E_{i,j}$. The ticked tails obtained in that way, are then extended by Procedure 1 recursively and occurrences are reported, if any. After the Procedure 1 ends, the KMP algorithm resumes and testing of the pattern starts at the beginning of the seed S_{i+1} .

5 Analysis

In this section, we discuss the correctness of the algorithm and analyse its space and time complexity.

Procedure 1: Procedure to extend ticked tails in a given \mathcal{TT}_i and reporting the occurrences found, if any.

input : A boolean array \mathcal{TT}_i of size m indicating ticked tails to be extended.
output: Reporting the found occurrences and preparing \mathcal{TT}_{i+1} for the next recursive call.

$isNonEmpty \leftarrow false$;
forall t **in** \mathcal{TT}_i **which are ticked** **do**
 $l_s \leftarrow |LCA(P[t+1..m], S_{i+1}[1..|S_{i+1}|])|$;
 if $(l_s + t) > m$ **then** // Pattern ends
 Report the occurrence;
 else if $l_s = |S_{i+1}|$ **then** // S_{i+1} occurs here
 $e \leftarrow t + |S_{i+1}|$;
 forall $E_{i+1,j}$ **in** ξ_{i+1} **do**
 $l_e \leftarrow |LCA(P[e..m], E_{i+1,j}[1..|E_{i+1,j}|])|$;
 if $(l_e + e) > m$ **then** // Pattern ends
 Report the occurrence (if not reported already);
 else if $l_e = |E_{i+1,j}|$ **then** // $E_{i+1,j}$ occurs here
 Mark $e + |E_{i+1,j}| - 1$ in \mathcal{TT}_{i+1} ;
 $isNonEmpty \leftarrow true$;
if $isNonEmpty$ **then**
 Extend(\mathcal{TT}_{i+1});

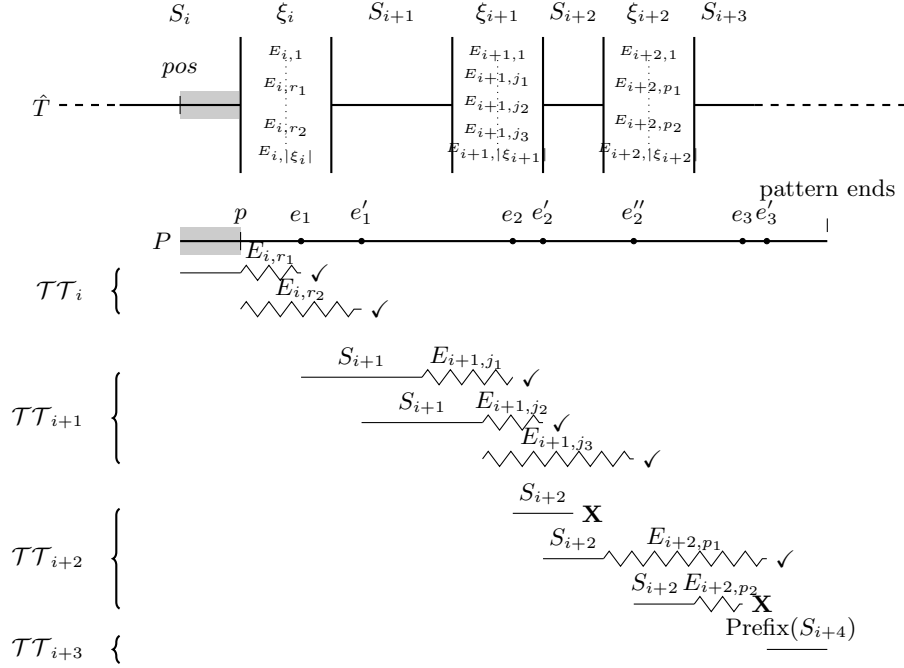


Fig. 1. An illustration of how the algorithm worked as described in Type 1. Strings in elastic-degenerate symbols have been shown as zigzag, while solid lines depict the seeds. Symbol **X** denotes that this path could not be extended further while the symbol \checkmark represents a ticked tail.

5.1 Correctness

Correctness of the presented algorithm is straightforward as every position of the text is being tested for an occurrence exhaustively. While the occurrences corresponding to the Cases 1 and 3(a) are covered by Type 1, Type 2 investigates all the occurrences associated with Case 2 and Case 3(b) - thus all the occurrences of P in \hat{T} are reported.

5.2 Space Complexity

The space needed by both, the failure-function and ticked tails array, is $\mathcal{O}(m)$. Each suffix tree \mathcal{ST}_{S_i} uses $\mathcal{O}(m + |S_i|)$ and \mathcal{ST}_{ξ_i} takes $\mathcal{O}(m + \sum_{i=1}^{k-1} \sum_{j=1}^{|\xi_i|} |E_{i,j}|)$ space, leading to the total space occupied by the tree to be $\mathcal{O}(km + N)$. Thus, assuming k to be constant, the solution only needs the space that is linear in the input size.

An important fact that can be exploited to make the algorithm further space-efficient is that all the suffix trees are not required in the memory at the same time. Once the start-position crosses past a seed or an elastic-degenerate symbol, their corresponding trees are no longer needed and can be discarded.

5.3 Time complexity

Time taken by the preprocessing stage is $\mathcal{O}(km + N)$ as the failure function can be computed in $\mathcal{O}(m)$ time and construction of all the suffix trees (along with their preprocessing required to answer LCA queries in constant time) can be done in $\mathcal{O}(km + N)$ time.

The search stage uses the KMP algorithm over each seed and each string of every elastic-degenerate symbol in the text, to report the occurrences for Case 1 and Case 2 and to search the beginning of the occurrence for Case 3. Thus, overall the time consumed by the KMP algorithm is $\mathcal{O}(\sum_{i=1}^k |S_i| + \sum_{i=1}^{k-1} \sum_{j=1}^{|\xi_i|} |E_{i,j}|)$ (i.e. $\mathcal{O}(N)$).

Procedure 1 can be analysed as follows: Intuitively, for every ticked position in the pattern (which can at most be m), LCP is calculated (in constant time) to find whether the corresponding seed occurs at the ticked position or not; a found such occurrence is then tried to extend by computing LCP with each of the strings in the following elastic-degenerate symbol. If α is the largest number of strings in any elastic-degenerate symbol of the text, this extension-step for each ticked position will be carried out at most α times. More specifically, the outer loop of the procedure runs m times and the inner one takes $\mathcal{O}(\alpha)$ time, as each LCA query takes constant time. Thus, each recursive call requires $\mathcal{O}(m\alpha)$ time. The number of recursive calls depends on the number of the elastic-degenerate symbols spanned by the occurrence of P being tested. In other words, if an occurrence spans across i elastic-degenerate symbols, there will be i recursive

calls to the procedure. If γ is the maximum such i , Procedure 1 executes in $\mathcal{O}(m\alpha\gamma)$ time (in total) for each start-position.

Initial ticking of the tails in Type 1 needs $\mathcal{O}(\alpha)$ time. For Type 2, initial ticking is done by KMP algorithm (already accounted above). In the worst case, Procedure 1 will be called from each of the n positions of the text, leading to an overall time-complexity of the algorithm to be $\mathcal{O}(nm\alpha\gamma + N)$ (as $k \leq n$). In real data, α and γ are mostly very small constants. Therefore, the algorithm is expected to work really efficiently in practice.

6 Conclusion

Motivated by the applications in genomics, we extended the notion of gapped strings to elastic-degenerate strings in this paper. We presented an efficient algorithm for the pattern matching problem, given a (solid) pattern and an elastic-degenerate text, running in $\mathcal{O}(N + \alpha\gamma nm)$ time; where m is the length of the given pattern; n and N are the length and total size of the given elastic-degenerate text, respectively; α and γ are small constants, respectively representing the maximum number of strings in any elastic-degenerate symbol of the text and the largest number of elastic-degenerate symbols spanned by any occurrence of the pattern in the text. Note that α and γ are so small in real-world applications that the algorithm is expected to work very efficiently in practice. The space used by the algorithm is linear in the size of the input for a constant number of elastic-degenerate symbols in the text.

It is to be noted that the presented algorithm can easily be adapted for a conserved (simple) degenerate pattern by using the algorithm given in [4] for conserved degenerate pattern matching. An interesting further direction is to conduct large-scale experiments, specifically in the context of studying inter-species genetic-variations¹. Furthermore, other domains that involve web-mining applications may find the presented solution interesting and beneficial.

References

1. Amir, A., Farach, M., Galil, Z., Giancarlo, R., Park, K.: Dynamic dictionary matching. *Journal of Computer and System Sciences* 49(2), 208 – 222 (1994), <http://www.sciencedirect.com/science/article/pii/S0022000005800479>
2. Church, D.M., Schneider, V.A., Steinberg, K.M., Schatz, M.C., Quinlan, A.R., Chin, C.S., Kitts, P.A., Aken, B., Marth, G.T., Hoffman, M.M., Herrero, J., Mendoza, M.L.Z., Durbin, R., Flicek, P.: Extending reference assembly models. *Genome Biology* 16(1), 13 (2015), <http://dx.doi.org/10.1186/s13059-015-0587-3>
3. Crochemore, M., Hancart, C., Lecroq, T.: *Algorithms on Strings*. Cambridge University Press (2007), 392 pages

¹ A proof-of-concept implementation of our algorithm can be accessed at <https://github.com/Ritu-Kundu/ElDeS>. Due to lack of space, experimental results are not included in the current version; they will be added in the full version of the paper.

4. Crochemore, M., Iliopoulos, C.S., Kundu, R., Mohamed, M., Vayani, F.: Linear algorithm for conservative degenerate pattern matching. *Engineering Applications of Artificial Intelligence* 51, 109 – 114 (2016), <http://www.sciencedirect.com/science/article/pii/S0952197616000130>, mining the Humanities: Technologies and Applications
5. Crochemore, M., Sagot, M.F.: *Motifs in Sequences: Localization and Extraction*, pp. 47–97. Marcel Dekker, New York (2004)
6. Diltthey, A., Cox, C., Iqbal, Z., Nelson, M.R., McVean, G.: Improved genome inference in the mhc using a population reference graph. *Nat Genet* 47(6), 682–688 (Jun 2015), <http://dx.doi.org/10.1038/ng.3257>, technical Report
7. Gusfield, D.: *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY, USA (1997)
8. Harel, H.T., Tarjan, R.E.: Fast algorithms for finding nearest common ancestors. *SIAM J. Comput.* 13(2), 338–355 (1984)
9. Huang, L., Popic, V., Batzoglu, S.: Short read alignment with populations of genomes. *Bioinformatics* 29(13), i361–i370 (2013), <http://bioinformatics.oxfordjournals.org/content/29/13/i361.abstract>
10. Knuth, D.E., James H. Morris, J., Pratt, V.R.: Fast pattern matching in strings. *SIAM Journal on Computing* 6(2), 323–350 (1977), <http://dx.doi.org/10.1137/0206024>
11. Liu, Y., Koyutürk, M., Maxwell, S., Xiang, M., Veigl, M., Cooper, R.S., Tayo, B.O., Li, L., LaFramboise, T., Wang, Z., Zhu, X., Chance, M.R.: Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. *BMC Genomics* 15(1), 685 (2014), <http://dx.doi.org/10.1186/1471-2164-15-685>
12. Maciuca, S., del Ojo Elias, C., McVean, G., Iqbal, Z.: A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference, pp. 222–233. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-43681-4_18
13. McCreight, E.M.: A space-economical suffix tree construction algorithm. *Journal of the ACM (JACM)* 23(2), 262–272 (1976)
14. Pissis, S.P.: Motex-ii: structured motif extraction from large-scale datasets. *BMC Bioinformatics* 15(1), 235 (2014), <http://dx.doi.org/10.1186/1471-2105-15-235>
15. Rahman, M.S., Iliopoulos, C.S., Lee, I., Mohamed, M., Smyth, W.F.: *Computing and Combinatorics: 12th Annual International Conference, COCOON 2006, Taipei, Taiwan, August 15-18, 2006. Proceedings*, chap. Finding Patterns with Variable Length Gaps or Don’t Cares, pp. 146–155. Springer Berlin Heidelberg, Berlin, Heidelberg (2006), http://dx.doi.org/10.1007/11809678_17
16. Schieber, B., Vishkin, U.: On finding lowest common ancestors: Simplification and parallelization. *SIAM J. Comput.* 17(6), 1253–1262 (Dec 1988), <http://dx.doi.org/10.1137/0217079>
17. Ukkonen, E.: On-line construction of suffix trees. *Algorithmica* 14(3), 249–260 (1995)
18. Weiner, P.: Linear pattern matching algorithms. In: *Proceedings of the 14th IEEE Annual Symposium on Switching and Automata Theory*. pp. 1–11. Institute of Electrical Electronics Engineer (1973)